Data Source: <https://www.kaggle.com/mylesoneill/world-university-rankings?select=timesData.csv>

**Introduction:**

The notion of conventional education being the conduit for success is something that has become a contentious topic in recent times. Traditionally, it was tacitly acknowledged that there was a linear relationship of sorts between the level of education one receives and their future success. The educational paradigm has continued to shift towards education being the most accurate measure of competency, especially in STEM fields. Jobs that once accepted high school degrees, are now requesting college degrees as a minimum with a preference for a masters. This is further compounded by the fact that those with advanced degrees make hundreds of thousands more over the course of their lifetimes than those without.

With such compelling data, it would appear as an intentional indictment of personal failure to forego the traditional educational system. Spanning the various socioeconomic classes, the utility of education varies. Amongst the poor, education is seen as a beacon of hope to elevate past one’s current conditions. The proverbial key of sorts to enter a higher social class and create a legacy. For the rich, education is perceived as more of a retention mechanism allowing the preservation of esteem to a family name or brand hence the aptly named term “legacy” at ivy league universities.

Despite all this, an injurious faction has materialized that is comprised of paradigm breaking entrepreneurs and jaded debt laden millennials. Contrarily, and equally compelling to the previous point, its common knowledge that the average university graduate carries tens of thousands of dollars of student debt. The total national student loan debt is in the trillions. The latter information attenuates the former however based on personal proclivities, the risk-reward could be tantamount.

With the aforementioned information taken into consideration, confusion is inevitable. Univariate analysis rarely paints a nuanced enough picture for decisive action. For this reason, the goal will be to predict accurately the score of a school given its ranking in the categories held in the dataset. Additionally, discovering which factors consistently hold importance in favorable ranking for an academic institution. This investigation will hopefully provide some reciprocal illumination for school selection and academic stakeholder satisfaction.

**About the data:**

The data set, World University Rankings, was acquired from Kaggle. The data set is an amalgam of the Times Higher Education World University Ranking, The Academic Ranking of World Universities, and Center for World University Rankings. There are 1024 observations and a total of 14 variables. The description of the variables is as follows:

1. “world\_rank” (Ordinal) – World rank for university
2. “institution” (Nominal) – Name of university
3. “country” (Nominal) – Country of each university
4. “national\_rank” (Ordinal) – Rank of university within its country
5. “quality\_of\_education” (Ordinal) – Rank for quality of education
6. “alumni\_employment” (Ordinal) – Rank for alumni employment
7. “quality\_of\_faculty” (Ordinal) – Rank for quality of faculty
8. “publications” (Ordinal) – Rank for publications
9. “influence” (Ordinal) – Rank for influence
10. “citations” (Numerical) – Number of students at the university
11. “broad\_impact” (Ordinal) – Rank for broad impact
12. “patents” (Ordinal) – Rank for patents
13. “score” (Numerical) – Total score, used for determining world rank
14. “year” – Year of ranking 2012-2015

****

The data preprocessing for this data set was minimal. The NAs in the dataset were omitted cleaning primarily the “broad\_impact” column of the missing 9% of values. Additionally, in preparation for the models, the country of origin and ranking columns were removed as well.



For the purposes of exploratory data analysis, four total columns were removed. The “institution” and “country” were removed as they were categorical variables and would be unhelpful in this instance. The following two, “broad\_impact” and “year” were also deemed temporarily irrelevant for the task. A plotted correlation matrix was exceedingly useful in visualizing the relationships between variables. The data points in many variable relationships were well distributed implying that they are likely all independent of each other and are not subject to collinearity. There are three exceptions. Influence, publications, and citations appear to have a cone-shaped relationship that goes from narrow to wide potentially hinting at a degree of multicollinearity.

The variable that will be predicted in the models is, score. Examining the variable relationships with score, no obvious patterns or trends appear in the data points. This independence is favorable in providing untainted outcomes in future predictions. The independence will provide diversity to the models and give less biased outcomes in their selection processes.

A picture containing diagram

Description automatically generated

The above image shows a correlation matrix of relevant variables in R.

**Models:**

The first model used for classification is RandomForest model. RandomForest is by virtue a more powerful and robust version of the decision tree classifier. The “forest” designation was earned from the fact that the forest is a collection of independent decision trees working in ensemble fashion. The independent decision trees output a class related decision that is put forth as a vote. Once this process is complete, the class votes are collected and like any good democracy, a final decision is made based on the majority. Low correlation is essential for the efficacy of the model. Low correlation implies high diversity amongst the trees which yields an unbiased decision. The RandomForest function utilizes the Breiman’s RandomForest algorithm and has the ability to accept many parameters. The four used for this RandomForest are:

1. x - XF\_TRAINING - a data frame or a matrix of predictors, or a formula describing the model to be fitted
2. y – YT\_TRAINING - A response vector. If a factor, classification is assumed, otherwise regression is assumed
3. maxnodes – 10 – Maximum number of terminal nodes trees in the forest can have
4. ntree - number of trees grown

Text

Description automatically generated

The above image shows the RandomForest() function with its default arguments.

The comparison model will be the neural net or neuralnet package. Neural networks as the name suggests, is an algorithm that is meant to be an abstraction mimicking how the human brain works. Multiple layers of networks are created from a series of nodes, often backbox, that are utilized for making predictions and classifications. The model typically features an input layer, a first level hidden layer, a second level or more hidden layers, and finally the output layer or predictions of the model. The connections to nodes, usually multiple, utilize logistic regression and will each be weighted differently. A neuron includes a bias term and activation function which comprise a sigmoid leading to the predicted probability. The four arguments used in this neuralnet function are:

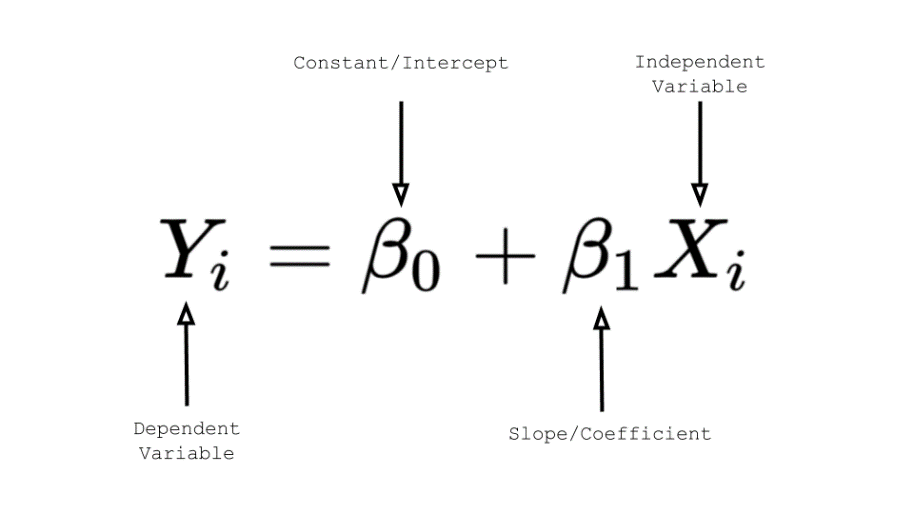
1. formula – a symbolic description of the model to be fitted
2. data - neural networkTRAINING – a dataframe containing the variables specified in the formula
3. Hidden – c(6,4,2) – a vector of integers specifying the number of hidden neurons in each layer
4. Linear.output – FALSE – logical. If act.fct should not be applied to the output neurons set linear output to TRUE, otherwise to FALSE

Text

Description automatically generated

The above image shows the neuralnet() function with its default arguments.

The final model used is a conventional linear regression. Linear regression is used for forecasting primarily. It assumes that there is one outcome variable (dependent) which can be modeled linearly from one or more input variable(s) (independent). Every input variable in this instance is a metric variable. In this analysis the university score is the dependent variable and the independent variables used to predict it are the quality of education, the alumni employment score, the quality of faculty rating, number of publications, influence, number of citations, the broader university impact, and the number of patents. The equation for linear regression can be show as:



The above image shows the equation for linear regression.

Y represents the dependent variable which is attempting to be predicted. Bo represents the constant or intercept. B1 are the slopes or coefficients of which the magnitudes of influence are measured to determine their effect on the dependent variable when multiplied with Xi the independent variable(s).

Graphical user interface, text

Description automatically generated

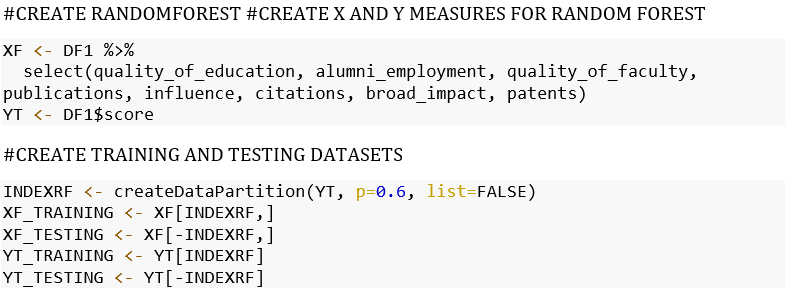
The above image shows the lm() function with its default arguments.

The two arguments used in this lm function are:

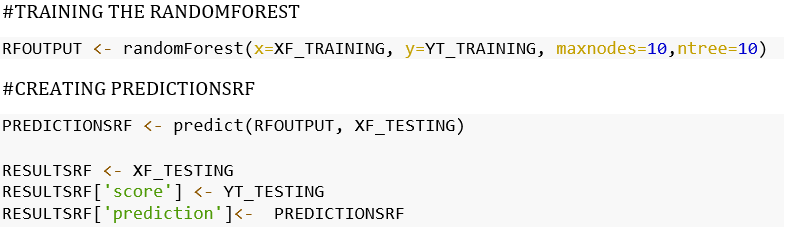
1. formula – a symbolic description of the model to be fitted
2. data – DF1 – a optional data frame, list, or environment containing the variables specified in the model

**Results:**

The first model utilized in this project is the RandomForest model. After the data is loaded and prepped, it is used in order to create two subsets. Those two subsets are the X and Y measures. The X measures are the variables/predictors and the Y measure is the goal. In our project the X measures are the variables that contribute to the final score, that score is the Y measure. We then further split the X and Y measures subsets into two subsets of each, a training and testing subset. The training subsets are made using 60 percent of the parent dataset (either the X or Y measures) and the training subsets are made with the remaining 40 percent.



The RandomForest model is trained by setting the X to the X measure training subset and Y to the Y measure training subset, a maxnodes set to 10, and the ntree set to 10. The trained RandomForest then uses the X measure testing dataset in order to make predictions. The predictions are compared against the Y measure testing dataset in order to compare the actual score to the predicted score.



We then took the statistical measurements of some of the most important measures of how accurate a model is. Those include the root mean-squared error and the r-squared values for the model. These measures help us better understand the accuracy of our RandomForest model. We also created a graph that can compare the accuracy of the models predicted vs real scores based on a single variable at a time. The graph below shows the impact of the patents variable on the real and predicted scores and helps show the accuracy of the trained RandomForest.



The above figure shows the root mean-squared error value found for the RandomForest model in R.



The above figure shows the R-squared value found for the RandomForest in R.

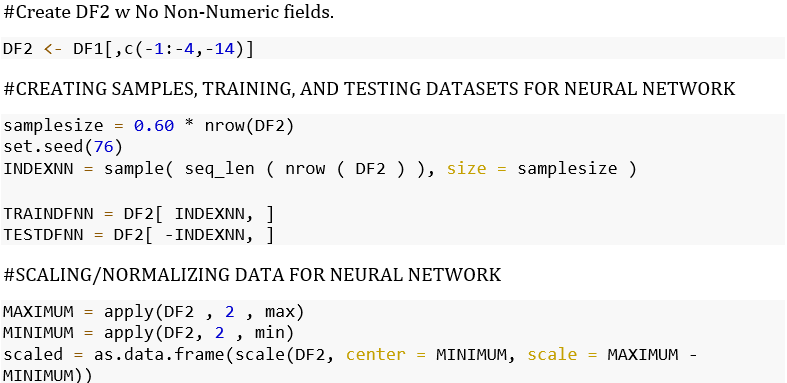
Chart, scatter chart

Description automatically generated

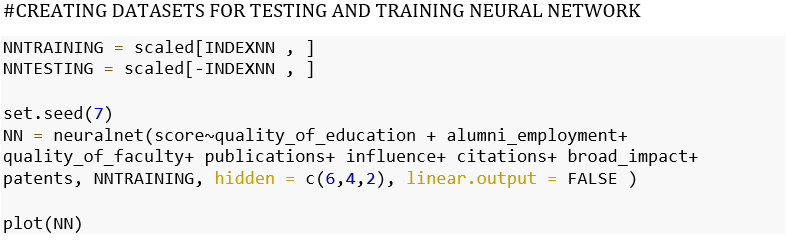
The above figure shows the patents value and its effect on the real and predicted scores for the RandomForest model in R.

The root mean-squared error tells us the square root of the average squared difference between the real and predicted values. The lower the root mean-squared error the better the model fits the dataset. And finally, the R-squared shows the amount of variance for the dependent variable that can be explained by the independent variables. The closer the R-squared value is to 1 the larger the effect of the variables have on the score. The R-squared found for the RandomForest model is high in this case, meaning the variables have a large effect on the score. The root mean-squared error can be interpreted by normalizing it as follows: NRMSE = RMSE / ( max value – min value ). The closer the normalized root mean-squared error is to 0 the better the fit of the model has to the data. In our case, the normalized root mean-squared error is 0.04272017 meaning the model has a very good fit and accuracy.

The next model ran in our project is the Neural Network model. The first step to setting up the neural network after loading the dataset and cleaning it was to create a subset of the dataset that removes the non-numeric fields. After the non-numeric fields are removed the dataset is split into two randomly sampled subsets, one with 60 percent of the data and the other with 40 percent. The larger of the two subsets is the training subset and smaller is the testing subset. Next, both subsets are normalized in order to help facilitate the neural network efficiency.



The normalized training subset is then used to train the neural network. We trained and tested the neural network in this project multiple times before deciding to have three hidden layers with a total of 12 nodes. Each layer has less nodes than the previous with the first layer having less nodes than the number of variables. This is a recommended best practice when training neural networks. We determined this combination as the most accurate while still minimizing model overfitting.



Diagram

Description automatically generated

The above figure shows the trained neural network model plotted in R. The accuracy of the model is roughly 93%.

The trained neural network then uses the normalized testing subset is to make predictions of the score. The predicted scores are compared to the real scores to determine the model’s accuracy. We then plotted the predicted scores against the real scores in order to more easily visualize the model’s accuracy. We also ran a nested loop with a set length of 65 in order to help validate the root mean-squared error for the model. As the length of the training set increased, the model’s root mean-squared error decreased, helping to validate the fitness of the model to the dataset.

Chart, scatter chart

Description automatically generated

The above figure shows the predicted vs real scores from the neural network plotted in R.

Chart, box and whisker chart

Description automatically generated

The above figure is a boxplot of the RMSE after the nested loop in R.

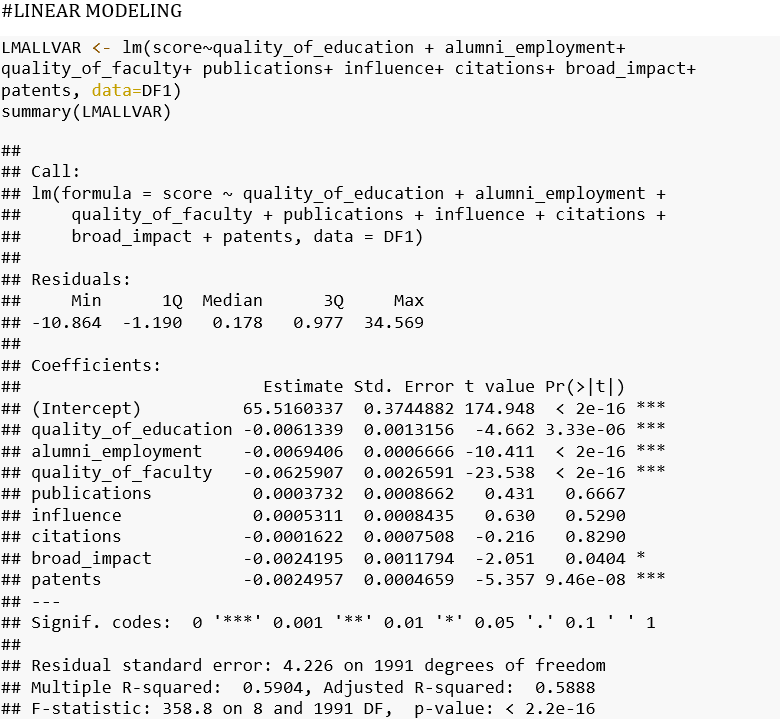
Chart, line chart

Description automatically generated

The above figure shows the median RMSE over training set time plotted in R. As the training set length increases the median RMSE decreases, this helps to validate the neural network model for this project.

The determined root mean-squared error was 0.766921 for the neural network. The normalized root mean-squared error was 0.01369991, using the normalized function previously mentioned during the RandomForest model results. This tells us that the model fitness is very good. In fact, the neural network NRMSE is even lower than the RandomForest model’s value. This indicates that the neural network model works more accurately and the model fits the dataset better as well.

After running these two models we wanted to identify the most significant variables. By finding the significant variables we can help determine the areas to focus on when institutions want to improve their global rankings. In order to do so, we decided to run a linear model after running the neural network.



The above figure shows the output of the linear model ran in R.

We ran the linear model and found that the model was significant, and that there were five significant variables besides the intercept. Out of the five significant variables, four were strongly significant and one was slightly significant. The four strongly significant variables were quality\_of\_education, alumni\_employment, quality\_of\_faculty, and patents. The one slightly significant variable was broad\_impact.

**Conclusion:**

For this project, the world university rankings dataset was used. First the dataset was loaded and cleaned. After the data was prepped, it was segmented for model use. For the RandomForest, we created four subsets, two training and two testing subsets. One of each for the X measure and one of each for the Y measure. For the Neural Network, we removed all non-numeric variables from the original dataset prior to creating two subsets, one training and one testing. After preparing the data for each model, we trained the models on the training subsets respectively. The trained models were used to make predictions on their corresponding testing subsets. Those predictions were compared to the actual scores to determine model accuracy. Validation measures were taken with each model to find its fit to the dataset.

Both models were carefully tuned so their accuracies and model fit were substantial while avoiding overfitting. The RandomForest model achieved a NRMSE value of 0.04272017 while the Neural Network achieved a NRMSE value of 0.01369991. Ultimately, it was determined the Neural Network model fit the dataset better and achieved higher accuracy. In addition to accuracy and fitness of the former, a linear model was run to further determine variables with the most impact on improving university ranking scores. The four variables with the most significant impact on the score were quality\_of\_education, alumni\_employment, quality\_of\_faculty, and patents. Improving these variables will in turn improve university score as effectively as possible. These tuned models are capable of accurately predicting a universty’s global ranking score based on these specific variables.